

Mineração de Dados em Biologia Molecular

Bases de Dados

André C. P. L. F. de Carvalho
Monitor: Valéria Carvalho



Aula de hoje

- Motivação
- SGBD
- Bases (bancos) de dados
- Desenvolvimento de sistemas de bases de dados
 - Modelo relacional
- Bases de dados biológicos

16/08/2012

André de Carvalho - ICMC/USP

2

Motivação

- Computadores são usados para manipular e armazenar dados
- Poucos dados: arquivos simples
- Grandes volumes de dados?

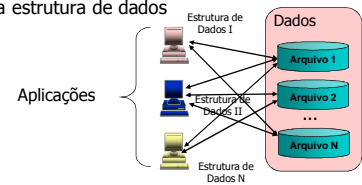
16/08/2012

André de Carvalho - ICMC/USP

3

Motivação

- Primeira opção: Sistemas de informação baseados em gerenciamento de arquivos
 - Rotinas específicas para tarefas específicas
 - Dados armazenados em disco, usando uma determinada estrutura de dados



16/08/2012

André de Carvalho - ICMC/USP

4

Motivação

- Problemas:
 - Redundância e inconsistência de dados
 - Múltiplos formatos de arquivos
 - Duplicação de informações em arquivos diferentes
 - Dificuldade de acesso aos/manipulação dos dados
 - Novo programa precisa ser escrito para realizar cada nova tarefa
 - Problemas de integridade
 - Restrições de integridade ficam escondidas no código, ao invés de explicitamente indicadas
 - Ex.: saldo da conta > 0

16/08/2012

André de Carvalho - ICMC/USP

5

Motivação

- Problemas
 - Falhas nas atualizações
 - Podem deixar a base de dados (BD) em um estado inconsistente, com atualizações apenas parciais
 - Ex.: transferência de recursos de uma conta para outras deve ou ser completa ou não ocorrer
 - Acesso concorrente por múltiplos usuários
 - Acesso concorrente sem controle pode levar a inconsistências
 - Dois usuários consultando e atualizando um arquivo ao mesmo tempo
 - Segurança
 - Difícil prover acesso a apenas parte dos dados

16/08/2012

André de Carvalho - ICMC/USP

6

Bases de dados relacionais

- Coleção de tabelas
 - Colunas representam os atributos dos dados
 - Todos os dados em uma coluna devem ser do mesmo tipo
 - Cada registro é armazenado em uma linha

Nome da tabela → **Pessoal**

Linha ou registro →

Nome	Endereco	Idade
José	Rua Sol 10	23
Maria	Av. La 23	18
Luiz	Rua Azul 20	57

Coluna ou campo →

10/06/07

7

Exemplo

- Dados de pacientes, amostras e seqüências
 - Um paciente tem uma ou mais amostras
 - Cada amostra tem uma ou mais seqüências

Código	Sexo	Data de nascimento	País	Código da amostra	Região do corpo	Data de coleta	Código da seqüência	Região	Formato FASTA
1	M	1/1/1970	Brasil	1	figado	10/10/2005	1111	ENV	'atgctgactgtctaccttggaaatcga'
1	M	1/1/1970	Brasil	1	figado	10/10/2005	1112	POT	'atgctgactgtctaccttggaaactaaa'
1	M	1/1/1970	Brasil	1	figado	10/10/2005	1113	Null	'aaatcga...tacttacttaccataactaaa'
1	M	1/1/1970	Brasil	2	baço	5/5/2005	2222	Null	'accataactaaa accataactaaa'
1	M	1/1/1970	Brasil	2	baço	5/5/2005	2223	POT	'tacttacttacttttacttaccataactaaa'
1	M	1/1/1970	Brasil	2	baço	5/5/2005	2223	POT	'tacttacttacttttacttaccataactaaa'
2040	F	2/2/1980	Angola	5	pulmão	2/2/2006	3333	ENV	'atgctgacta...tacttacttaccataactaaa'
2040	F	2/2/1980	Angola	5	pulmão	2/2/2006	3334	Null	'aaatcga...tacttacttaccataactaaa'
2040	M	2/2/1980	Angola	5	pulmão	2/2/2006	3335	POT	'aaatcga...tacttacttaccataactaaa'

Exemplo

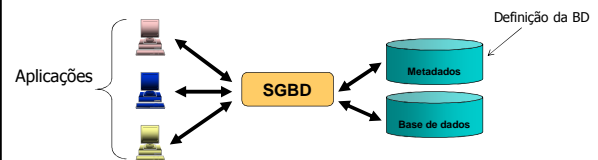
Replicado 6 vezes Possui 2 amostras Cada amostra tem 3 seqüências

Código	Sexo	Data de nascimento	País	Código da amostra	Região do corpo	Data de coleta	Código da seqüência	Região	Formato FASTA
1	M	1/1/1970	Brasil	1	figado	10/10/2005	1111	ENV	'atgctgactgtctaccttggaaatcga'
1	M	1/1/1970	Brasil	1	figado	10/10/2005	1112	POT	'atgctgactgtctaccttggaaactaaa'
1	M	1/1/1970	Brasil	1	figado	10/10/2005	1113	Null	'aaatcga...tacttacttaccataactaaa'
1	M	1/1/1970	Brasil	2	baço	5/5/2005	2222	Null	'accataactaaa accataactaaa'
1	M	1/1/1970	Brasil	2	baço	5/5/2005	2223	POT	'tacttacttacttttacttaccataactaaa'
1	M	1/1/1970	Brasil	2	baço	5/5/2005	2223	POT	'tacttacttacttttacttaccataactaaa'
2040	F	2/2/1980	Angola	5	pulmão	2/2/2006	3333	ENV	'atgctgacta...tacttacttaccataactaaa'
2040	F	2/2/1980	Angola	5	pulmão	2/2/2006	3334	Null	'aaatcga...tacttacttaccataactaaa'
2040	M	2/2/1980	Angola	5	pulmão	2/2/2006	3335	POT	'aaatcga...tacttacttaccataactaaa'

- Estrutura não otimizada
- Usa muito espaço em disco
- Cada atualização de paciente ou amostra deve ser propagada em todas as linhas do paciente ou amostra

Alternativa

- Uso de um sistema intermediário que torna os programas independentes da estrutura de dados
 - Sistema de gerenciamento de banco de dados (SGBD)



16/08/2012

André de Carvalho - ICMC/USP

10

Vantagens de um SGBD

- Armazenamento persistente de dados e estruturas de dados
- Independência dos dados
- Consistência dos dados
- Acesso compartilhado à informação (multi-usuário e concorrente)
- Distribuição de informações
- Reduz complexidade das aplicações
- Controle de acesso aos dados
 - Segurança
 - Facilita Backup

16/08/2012

André de Carvalho - ICMC/USP

11

Sistemas de arquivo tradicionais x SGBD

Arquivos tradicionais	SGBD
Definições dos dados é parte do código dos programas	Meta-dados
Dados e aplicação são dependentes	Dados e aplicação são independentes
Dados representados no nível físico	Representação conceitual
Cada módulo implementa uma visão dos dados	Múltiplas visões dos dados

16/08/2012

André de Carvalho - ICMC/USP

12

Base de Dados

- Definição:
 - Coleção de dados logicamente relacionados que tem algum propósito associado
- Projetada, construída e preenchida com dados para satisfazer um propósito ou público específico
- Representa algum aspecto do mundo real
 - Mini-mundo

16/08/2012

André de Carvalho - ICMC/USP

13

SGBD

- Tem funções para **definir (incluir), recuperar, excluir e modificar** dados em uma BD

16/08/2012

André de Carvalho - ICMC/USP

14

SGBD

- Informação no sistema de BD deve ter uma estrutura
- Descrição da estrutura = esquema
 - Mantida na BD como metadados
- Conceitos e associações do mundo real devem ser capturados do mini-mundo e armazenados como metadados do SGBD
 - Exemplo:
 - Mini-mundo: universidade
 - Conceitos: cursos, disciplinas, aulas, alunos
 - Associações: cursos tem disciplinas, alunos se matriculam em cursos

16/08/2012

André de Carvalho - ICMC/USP

15

Exemplo

Domínio(sexo) = carácter indicando o sexo: M (masculino) e F (feminino)

Domínio(país) = cadeia de caracteres com o nome do país de origem do paciente

Atributo

codPaciente	sexo	nascimento	país
0001	M	1/1/1970	Brasil
1119	M	1/4/1970	Brasil
1209	F	16/10/1970	Chile
0002	M	21/1/1997	EUA
1987	F	15/1/1979	Brasil
1111	F	1/1/1980	Angola
2040	F	2/2/1980	Angola

Valor

Tupla

16/08/2012

André de Carvalho - ICMC/USP

16

SQL

- Structure Query language
- É uma linguagem de computador que segue o padrão ANSI
- SQL permite:
 - Consultar uma BD relacional
 - Recuperar dados de uma BD
 - Inserir novos registros em uma BD
 - Deletar registros de uma BD
 - Atualizar registros em uma BD

10/06/07

17

SGBDs que usam SQL

- Produtos comerciais
 - Microsoft ACCESS (Microsoft Office)
 - Microsoft SQLserver
 - Oracle
- Freeware
 - **MySQL**
 - PostgreSQL
 - MiniSQL

16/08/2012

André de Carvalho - ICMC/USP

18

Bases de dados biológicas

- Gerais
 - Sequências de DNA, funções de proteínas, estruturas 3-dimensionais de proteínas, ...
- Especializados
 - EST, STS, SNP, RNA, genomas, famílias de proteína, pathways, dados de microarray, ...

16/08/2012

André de Carvalho - ICMC/USP

19

BD de genomas

- Toda sequência de genoma publicada deve ser disponibilizada em uma BD pública
 - Membros do International Nucleotide Sequence Database Collaboration são os principais repositórios
 - Consórcio formado por 3 grandes BDs
 - [EMBL](#) (European Molecular Biology Laboratory nucleotide sequence database at [EBL](#), Hinxton, UK)
 - [GenBank](#) (at National Center for Biotechnology information, [NCBI](#), Bethesda, MD, USA)
 - [DDBJ](#) (DNA Data Bank Japan at [CIB](#), Mishima, Japan)

16/08/2012

André de Carvalho - ICMC/USP

20

Conclusão

- Motivação
- SGBD
- Bases de dados
- Desenvolvimento de Sistemas de BDs
 - Modelo Relacional
- BD Biológicos

16/08/2012

André de Carvalho - ICMC/USP

21

Agradecimento

- Prof Carlos Eduardo Ferreira, IME-USP

16/08/2012

André de Carvalho - ICMC/USP

22

Perguntas

André Ponce de Leon F de
Carvalho

23